Array Annotation

Doron Betel Weill Cornell Medical College March 23, 2012

HELP custom arrays were annotated by the following procedure:

1. Probe sequences were extracted from the corresponding `*.ndf` files and converted to fasta format.
2. Probes were aligned to the latest genome builds (mm9 or hg19) by BLAT using default parameters. The resulting `*.blat` files can be uploaded to the UCSC genome browser to view the genomic location of the probes and confirm the probeset annotation.
3. Probes with perfect genome matches were annotated to proximal genes from both strands (`*.annot` files). The annotation includes **all** genes with a transcription start sites (TSS) within 10kb bases from the probe (end), or to the gene with the closest TSS. Note that every probes is annotated separately to each strand.
4. Probe level information was combined to probeset annotation (`*.probesetannot` files). The probes were originally designed in groups of 5-6 covering CpG islands. In some cases one or more probes in a probeset are mapped to multiple genomic regions, which can also be on the same chromosome. In other cases the enitre probset is align to repeat regions and therefore appears in multiple loci. In all those cases, each genomic occurrence is annotated separately and will appear in separate lines in the annotation file. Each probeset annotation line is marked by a boolean flag `[True, False]` that indicates if the annotation was derived from probes overlapping the probeset (`True`) or outside the probeset (`False`).
5. Probeset locations were converted to `*.bed` files according to the specification in `*.ngd` files so they can be viewed with the probe alignments in the genome browser.
6. Probes that aligned to positions beyond their intended probeset design are listed in a separate file `data/*.nsp`. These probes may be removed from downstream analysis.
7. Probeset were annotated to CpG islands, provided in `*.cpg` files. CpG island locations were downloaded from UCSC genome browser for hg19 and mm9 and files are included in the data dir. Similarly to probe annotation, all CpG island sites within 10kb of probeset ends are recorded in the file.

The final `probsetannotation` file contain the following fields:

1. `Probe Id`
2. `Chromosome`: probeset chromosome annotation.
3. `Strand`: mapped strands. Some probesets contain probes from both strands.
4. `Probeset start`: Start position of the probeset determined by the left most probe. Note that genomic coordinates are always from left to right. Start position of '-' strand probesets are numbered with respect to '+' strand (the condition $start < end$ is always true)
5. `Overlap`: A boolean field with `[True, False]` values to indicate if this annotation corresponds to the probeset genomic location (as indicated in `*.ngd` file).
6. `Probeset end`: End position of probeset.
7. `Refseq id(s)`: Matched refseq id(s). All refseq that were annotated to **any** of the probes in the probeset according to the criteria outlined above. Multiple refseqs are comma separated (',') and the values in subsequent fields corresponds to these refseq genes.
8. `Refseq chr`: Refseq chromosome location.
9. `Refseq strand`: Refseq strand orientation.
10. `Gene symbol(s)`: Gene symbols corresponding to the refseq ids.
11. `Distance to TSS`: Distance to the refseq TSS.